

CSE 422 Autumn 26 HW 3

Note: For all problems, you will be expected to explain your reasoning. Providing an expression without reasoning will not receive any partial credit.

Problem 1: Lower bounds for hypothesis selection

Recall that we showed the following in class. Let $\varepsilon > 0$, and let D be a distribution over $\mathbb{R}^d$ and $\mathcal{A} = \{f_1, \dots, f_k\}$ be a set of classifiers so that

$$\Pr_{X \sim D} [f_i(X) \neq f_j(X)] \geq \varepsilon \quad \text{for all } i \neq j. \quad (1)$$

Then, given n independent samples $X_1, \dots, X_n \sim D$ and labels $y_i = f(X_i)$ for some $f \in \mathcal{A}$, where $n \geq \Omega(\frac{\log n}{\varepsilon})$, the empirical risk minimizer uniquely identifies f with probability at least 0.99. The goal of this problem is to show that the $\log n$ dependence is necessary.

Specifically, you should come up with a distribution D , and a set of k classifiers that satisfy Equation 1, so that for some universal constant c , if you take less than $c \cdot \log k$ samples, no algorithm can uniquely recover the true classifier with probability $\geq 2/3$. Think carefully about how you can reason that there is be no classifier!

Problem 2: VC dimension

We alluded to the idea of Vapnik-Chernonenkis (VC) theory in class as a way to get generalization guarantees for infinite families of classifiers. In this problem, we'll define the key quantity relevant for this theory, namely, *VC dimension*.

Let $\mathcal{A}$ be a set of classifiers, i.e. a family of functions $f : \mathbb{R}^d \rightarrow \{0, 1\}$. Given a set of points $S \subset \mathbb{R}^d$, we say that $\mathcal{A}$ *shatters* S if for every subset $T \subseteq S$, there exists a $f \in \mathcal{A}$ so that $f(x) = 1$ for all $x \in T$, and $f(x) = 0$ for all $x \in S \setminus T$.

The *VC dimension* of a set of classifiers $\mathcal{A}$ is the size of the largest set S that can be shattered by $\mathcal{A}$. You will not need to prove this, but it turns out the sample complexity of learning a class of binary classifiers with VC dimension D can be learned to PAC error ε with probability $1 - \delta$ with n samples, as long as

$$n = \Theta\left(\frac{D + \log 1/\delta}{\varepsilon}\right).$$

This motivates our interest in bounding the VC dimension of family of classifiers!

Our goal in this problem is to establish the VC dimension of the class of linear classifiers that go through the origin. For a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, let $\mathbf{1}_f(x) = 1$ if $f(x) \geq 0$, and 0 otherwise, and let

$$\mathcal{A} = \left\{ \mathbf{1}_{\langle v, x \rangle}(x) : v \in \mathbb{R}^d \right\}.$$

(a) [7 points] Establish that there exists a set S of d points so that $\mathcal{A}$ shatters S .

(b) [7 points] Demonstrate that for *any* set S of $d + 1$ points, $\mathcal{A}$ *cannot* shatter S .

These two parts together imply that the VC dimension of linear classifiers is d , which shows that the sample complexity of learning them is linear in d .

(c) [6 points] For some natural number $k \geq 1$, let $\mathcal{R}_k$ denote the set of rational functions $p(x)/q(x)$ from $\mathbb{R}^d$ to $\mathbb{R}$ where both p and q are polynomials of degree at most k and q satisfies $q(x) > 0$ for all $x \in \mathbb{R}^d$, and let $\mathcal{A}_k = \{\mathbf{1}_f(x) : f \in \mathcal{R}_k\}$. Give an upper bound for the VC dimension of $\mathcal{A}_k$.

Problem 3: Regularization and measurement error

In this problem, you will see the benefits of regularization in a fairly generic setting. Suppose we have a linear regression problem with predictor matrix $X \in \mathbb{R}^{n \times d}$, unknown ground truth θ^* , and feedback

$$y = X\theta^* + \xi ,$$

where the coordinates of ξ are independent, have mean zero, and satisfy $\text{Var}[\xi_i] = \sigma^2$ for all $i = 1, \dots, n$. Assume that

$$X^\top X = \begin{pmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & s_d \end{pmatrix} ,$$

and for all regularization parameters $\lambda > 0$, define the ℓ_2 regularized solution

$$\theta_\lambda = \arg \min_{\theta} \left\{ \|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2 \right\} .$$

(a) [5 points] Derive a closed form solution for θ_λ , for all λ , in terms of X and y .

(b) [5 points] Using this solution, compute an expression for the expected bias from θ , and variance of every coordinate of θ_λ . How do these expressions change as we vary λ ?

(c) [5 points] For all $\lambda > 0$, derive a formula for the mean-squared error, in terms of the parameters s, θ^*, σ^2 , and λ :

$$\text{MSE} = \mathbb{E}[\|\theta_\lambda - \theta^*\|_2^2] .$$

(c) [5 points] Show that, for some settings of parameters s, θ^* , and σ^2 , there is some choice of λ that decreases the mean-squared error.